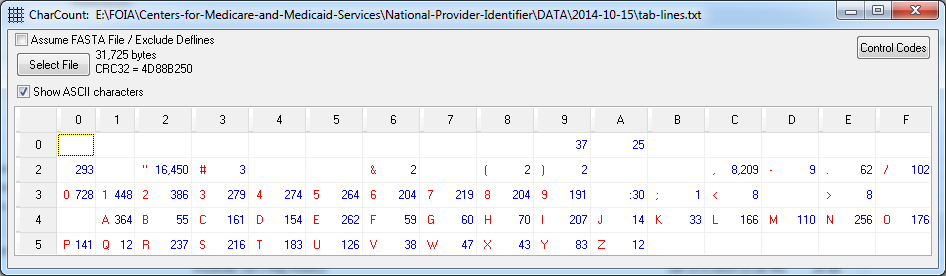


37 tabs (hex 09)

Write lines with tabs to *tab-lines.txt* for inspection

[112 efg localhost 2014-11-02 08:08:34 /mnt/hgfs/E/FOIA/Centers-for-Medicare-and-Medicaid-Services/National-Provider-Identifier/DATA/2014-10-15]

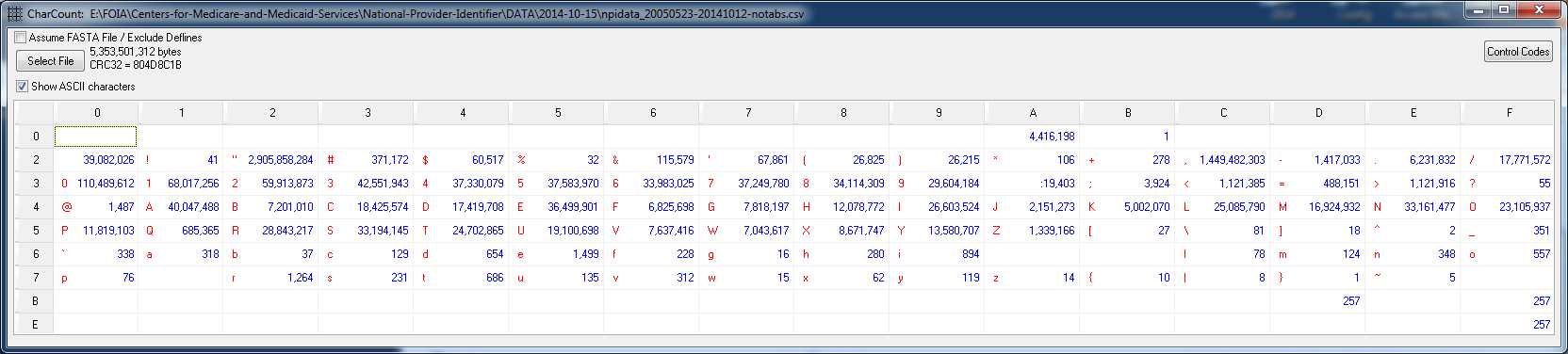
grep -n $'\x09' npidata\_20050523-20141012.csv > tab-lines.txt



Tabs in the 25 lines of *tab-lines.txt* show there will be no loss if tabs are converted to spaces.

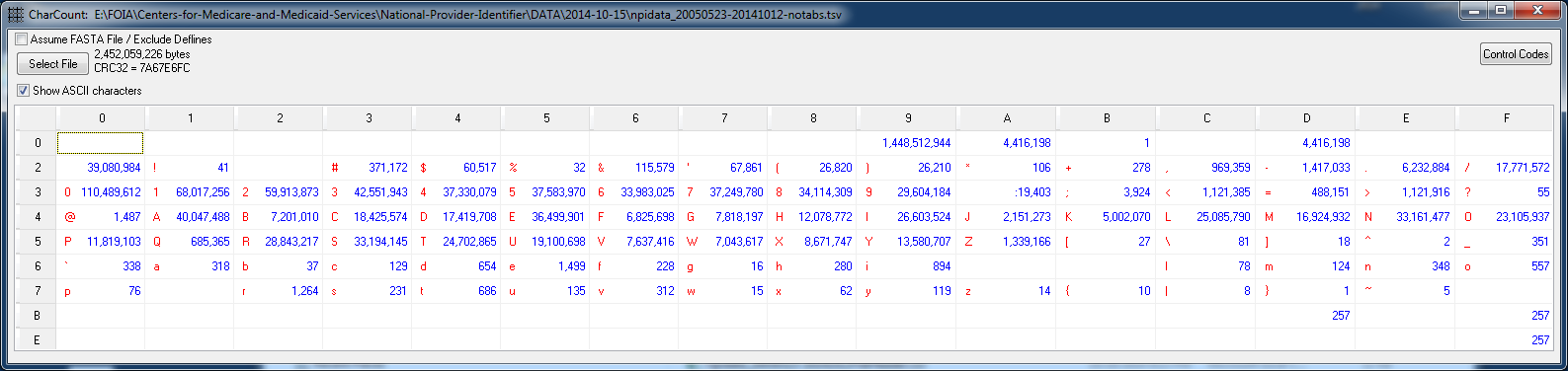
[113 efg localhost 2014-11-02 08:14:47 /mnt/hgfs/E/FOIA/Centers-for-Medicare-and-Medicaid-Services/National-Provider-Identifier/DATA/2014-10-15]

tr $'\x09' ' ' < npidata\_20050523-20141012.csv > npidata\_20050523-20141012-notabs.csv



Above is proof 37 tabs removed, and 37 spaces inserted.

After conversion of all tabs, we can re-write the file with a tab delimiter and remove all enclosing quotation marks (about 2 billion of them!)



The number of tabs above (1,448,512,944) does not make sense.

2,905,858,284 double quotes / 2 = 1,452,929,142 tabs.

National Plan and Provider Enumeration System (NPPES) Downloadable File  
<http://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/NationalProvIdentStand/DataDissemination.html>

Download National Provider Identifier (NPI) Downloadable File  
<http://nppes.viva-it.com/NPI_Files.html>

42 of 329 fields loaded into Access NPI table: 4,416,196 records

Summary of Access queries

**Organizations**

* 1,035,094 all
* 707,546 unique addresses (includes room, suite, box number, …)
* 13,379 MO unique addresses

**People**

* 3,310,313 all
* 1,976,415 unique addresses (includes room, suite, box number, …)
* 31,220 MO unique addresses